Statistics and Applications {ISSN 2452-7395 (online)}
Volume 18, No. 1, 2020 (New Series), pp 307–318

Change Point Modeling of Covid-19 Data in the United States

Sheng Zhang[1], Ziyue Xu[2] and Hanxiang Peng[1]
[1]*Indiana University-Purdue University Indianapolis*
[2]*University High School of Indiana*

Received: 06 July 2020; Revised: 26 July 2020; Accepted: 28 July 2020

Abstract

To simultaneously model the change point and the possibly nonlinear relationship in the Covid-19 data of the US, a continuous second-order *free* knot spline model was proposed. Using the least squares method, the change point of the daily new cases against the total confirmed cases up to the previous day was estimated to be 04 April 2020. Before the point, the daily new cases were proportional to the total cases with a ratio of 0.287, suggesting that each patient had 28.7% chance to infect another person every day. After the point, however, such ratio was no longer maintained and the daily new cases were decreasing slowly. At the individual state level, it was found that most states had change points. Before its change point for each state, the daily new cases were still proportional to the total cases. And all the ratios were about the same except for New York State in which the ratio was much higher (probably due to its high population density and heavy usage of public transportation). But after the points, different states had different patterns. One interesting observation was that the change point of one state was about 3 weeks lagged behind the state declaration of emergency. This might suggest that there was a lag period, which could help identify possible causes for the second wave. In the end, consistency and asymptotic normality of the estimates were briefly discussed where the criterion functions are continuous but not differentiable (irregular).

Key words: Asymptotic normality; Change point; Consistency; Covid-19; Free knot; Irregular criterion function.

1. Introduction

The first case of Novel Coronavirus disease 2019 (Covid-19) was reported in Wuhan, China on 17 November 2019. This disease was caused by SARS-CoV-2 virus, and in about 6 months, it has spread throughout the whole world, infected 15.5 million people, and killed more than 635,000 (https://covid19.who.int/). In the United States, there are 4 million confirmed cases, and 143,000 deaths by July 25th. Many states have ordered their residents to stay at home and keep social distancing to slowdown the rapid spread of the virus, so that the health care system will not be overwhelmed. The trend of daily new cases in the US appeared to be flattened in the early April. Here, we first fitted the data with the change point model (Bai, 1997; Julious, 2000) to identify the possible dates for the trend change.

The first case in US was reported on 21 January 2020 in Washington State. By the end of February, several more confirmed cases were recorded there. By the end of March, the number of confirmed cases quickly went up to about 6,000. On 29 February 2020, the Governor declared the state emergency. A few weeks later, the daily new cases stabled and slowly started

Corresponding Author: Hanxiang Peng
Email: hanxpeng@iupui.edu

decreasing. Similar patterns can also be observed in other states. By fitting the state data with a change point model, we found that the change point was correlated with the date when the state emergency was declared. Thus, we believed that one main possible cause for the change point could be the declaration of state emergency. Once people started to protect themselves more carefully, the effect of the protection would be noticeable after 2-3 weeks. Knowledge about this delay period would help us identify the causes if the trend changes again.

In this study, we used the data collected by New York Times Company. The data is stored at GitHub (http://github.com/nytimes/covid-19-data/blob/master/). It contains the number of cumulated cases at the county level, state level, and country level, starting from 21 January 2020. We downloaded the data up to 18 June 2020 for this study.

2. Change point model and data fitting procedure

First, we fitted the data at the county-level. Displayed in Fig. 1A is the plot of the number of daily new cases against the total number of cases up to the previous day. Noticeably, there is a change point between 27 March and 05 April 2020, around which an increasing relationship of the daily new cases against the total cases was progressed to decreasing. Specifically, at first, the number of daily new cases was drastically increasing with the total number of cases up the previous day. Then after some critical point, the increasing relationship turned to decreasing but at a slow rate. This seems to be no surprising. When Covid-19 broke out, a great number of people got infected within a short period of time. Meanwhile, measures such as social distancing and using of personal protective equipment were taken, the spreading was slowed down. Motivated by these plots, we chose to use a change point model to fit the data.

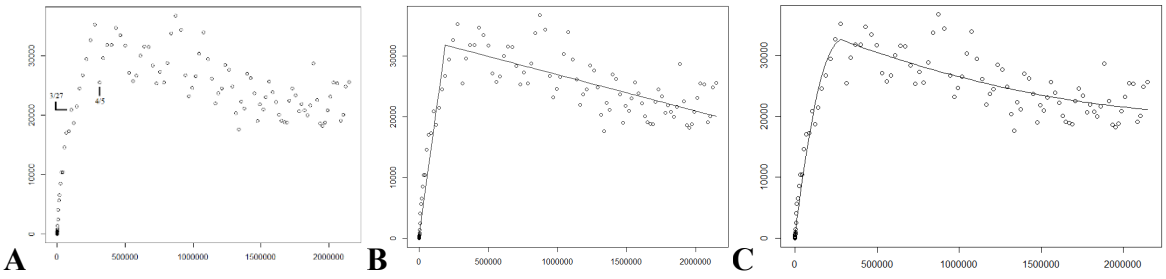


Figure 1: The scatter plot of data. The *y*-axis is the number of daily new cases, and the *x*-axis is the total number of cases up to the previous day. **A**: the scatter plot of the data; **B**: the scatter plot superimposed with the fitted linear model (1); **C**: the scatter plot superimposed with the fitted quadratic model (2).

In a linear change point model, the expected value $E(y_i)$ of the number y_i of daily new cases is expressed as a linear function of the total number x_i, of cases up to the previous day, i.e.,

$$E(y_i) = \beta_0 + \beta_1 x_i + \beta_2 (x_i - \delta)_+, \quad i = 1, \dots, n \tag{1}$$

where δ is an unknown change point, and $x_+ := \max(x, 0)$ is the positive part of x. Here β_0 is the intercept, which is expected to be very close to zero (there should be almost no new case if there is no confirmed cases), β_1 is the rate of infection before the change point, which can be interpreted as how many persons will be infected by each patient every day; β_2 can be interpreted as the effectiveness of the protective measures taken to stop the disease. Treating δ known, we estimate the parameters beta by the least squares method. To estimate the change

point δ, we can search all possible values of the change point and compare the corresponding sum of squared residuals (SSE). The estimator of change point is the one corresponding to the smallest SSE. The data in Fig. 1A was fitted to the linear model with one change point, with the daily new cases as the response and the total number of cases up to the previous day as the predictor. Listed in Table 1 are all the possible change points with corresponding SSE values.

Table 1: The possible change points and corresponding SSE from linear model

Date	δ	SSE$\times 10^{-9}$
2020-03-28	102835	1.34
2020-03-29	123730	1.23
2020-03-30	142406	1.14
2020-03-31	163873	1.08
2020-04-01	**188425**	**1.05**
2020-04-02	215176	1.08
2020-04-03	244636	1.17
2020-04-04	277279	1.34
2020-04-05	312519	1.61

From the Table 1, the estimated change point is 01 April 2020, which is consistent with our observation. The results of all other parameters are listed in Table 2. With this change point, the fitted equation is:

$$E(y) = 495 + 0.1662x - 0.1722(x - 188425)_+$$
$$= \begin{cases} 495 + 0.1662x, & x \le 188425, i.e., \text{before 01 April 2020} \\ 32942 - 0.006x, & x > 188425, i.e., \text{after 01 April 2020} \end{cases}$$

In this equation, $\hat{\beta}_0$ is not significantly different from zero, which was consistent with our intuition: β_0 should be very close to zero. $\hat{\beta}_1 = 0.1662$ indicated that before the change point, each patient had 16.6% chance to infect another healthy person every day. $\hat{\beta}_2 = -0.1772$ suggested that after the total number of confirmed cases reached to 188,425, the slope of the linear trend is $(\hat{\beta}_1 + \hat{\beta}_2) = -0.006$. This showed that the daily new cases were decreasing after 01 April 2020, but at a very slow rate.

Displayed in Fig. 1B is the scatter plot of the data superimposed with the fitted lines, using 01 April 2020 as the change point. The simple linear model fit data well, except that there are some noticeable non-linear features for both before and after the change point. This motivates us to fit the data with a continuous quadratic change point model:

$$E(y_i) = \beta_0 + \beta_1 x_i + \beta_{12} x_i^2 + \beta_2 (x_i - \delta)_+ + \beta_{22} (x_i - \delta)_+^2 \quad (2)$$

where β_1 is the initial rate when there is only a small amount of confirmed cases; β_{12} is the correction factor for the non-linear feature before the change point; β_2 and β_{22} indicate the effectiveness of the prevention measures after the change point. Our study exhibited in this model that the LSE is asymptotic normal. The estimation method is the same as described above, and the possible change points and their corresponding SSE are listed in Table 3.

In this model, the change date is 04 April 2020, and all the other estimated parameters are listed in Table 2. The fitted equation is:

$$E(y) = 290 + 0.2247x - 3.898 \times 10^{-7}x^2 - 0.01845(x - 277279)_+ + 3.919 \times 10^{-7}(x - 277279)_+^2$$

$$= \begin{cases} 290 + 0.2247x - 3.898 \times 10^{-7}x^2, & \text{before 04 April 2020} \\ 35531 - 0.01106x + 2.028 \times 10^{-9}x^2, & \text{after 04 April 2020} \end{cases}$$

The superimposed plot is shown in Fig. 1C. The quadratic model appeared to be a better fit to the data. To confirm this, we performed ANOVA test to test if the linear model is sufficient. The ANOVA test result is shown in Table 4, indicating that this full model is appropriate. Another question that arises is - should we still pick 01 April 2020 as the change point as suggested from the linear model? The ANOVA test result is listed in Table 4, from which our answer would be that 04 April 2020 should be the change point. Possibly the linear model is somewhat oversimplified, as it ignores the curve feature before and after the change point, which could lead to restrictions on selecting the change point due to its lack of flexibility. Thus, we would suggest that the change point for the US is 04 April 2020.

Table 2: The estimated coefficients from model (1)-(3)

Model (1): $E(y_i) = \beta_0 + \beta_1 x_i + \beta_2 (x_i - \delta)_+$

Estimator	Estimated value	Std Err	t^*	P ($t > t^*$)
$\hat{\beta}_0$	495	335	1.474	0.143
$\hat{\beta}_1$	0.1662	0.0037	44.533	0.000
$\hat{\beta}_2$	−0.1722	0.0041	−41.685	0.000

Model (2): $E(y_i) = \beta_0 + \beta_1 x_i + \beta_{12} x_i^2 + \beta_2 (x_i - \delta)_+ + \beta_{22} (x_i - \delta)_+^2$

Estimator	Estimated value	Std Err	t^*	P ($t > t^*$)
$\hat{\beta}_0$	290	323	0.898	0.371
$\hat{\beta}_1$	0.2247	0.00146	15.356	0.000
$\hat{\beta}_{12}$	-3.898×10^{-7}	5.56×10^{-8}	−7.01	0.000
$\hat{\beta}_2$	−0.01845	0.01798	−1.025	0.307
$\hat{\beta}_{22}$	3.919×10^{-7}	5.53×10^{-8}	7.091	0.000

Model (3): $E(y_i) = \beta_0 + \beta_1 x_i + \beta_{12} x_i^2 + \beta_2 (x_i - \delta)_+ + \beta_{22} (x_i - \delta)_+^2$ +weekly effect

The residual has AR (1) pattern

Estimator	Estimated value	Std Err	t^*	P ($t > t^*$)
$\hat{\beta}_0$	132	197	0.671	0.504
$\hat{\beta}_1$	0.2871	0.0161	17.825	0.000
$\hat{\beta}_{12}$	-5.143×10^{-7}	4.37×10^{-8}	−11.77	0.000
$\hat{\beta}_2$	-0.0128	0.011	−1.158	0.249
$\hat{\beta}_{22}$	5.17×10^{-7}	4.36×10^{-8}	11.87	0.000
Monday effect	−10130	1586	−6.385	0.000
Tuesday effect	−8960	1595	−5.618	0.000
Wednesday effect	−7848	1602	−4.9	0.000

Thursday effect	−4933	1611	−3.062	0.003
Friday effect	−3513	1557	−2.256	0.026
Saturday effect	−5379	1566	−3.44	0.001
Sunday effect	−10090	1577	−6.397	0.000

Table 3: The possible change points and corresponding SSE from the quadratic model

Date	Value of δ	SSE$\times 10^{-9}$
2020-03-28	102835	1.2719
2020-03-29	123730	1.1982
2020-03-30	142406	1.1266
2020-03-31	163873	1.0520
2020-04-01	188425	0.9893
2020-04-02	215176	0.9446
2020-04-03	244636	0.9190
2020-04-04	**277279**	**0.9111**
2020-04-05	312519	0.9112
2020-04-06	337984	0.9163
2020-04-07	367599	0.9285
2020-04-08	399388	0.9454

In Figure 1A, one notices that besides the trend, the variation of daily new cases exhibited/s strong weekly effect: during the weekend, the number was small, and during the middle of a week, the number was high. Here, the plot of the residual after 18 March 2020 is shown in Figure 2A. The plot indicated that there was an oscillation pattern. The auto-correlation function (ACF) plot of the residual is shown in Figure 2B. From the ACF plot, the weekly effect was apparent: the residual was highly positive correlated on 7 days and 14 days.

Table 4: The ANOVA test results

Full model: Quadratic model (2); reduced model: linear model (1)

Model	SSE	DF	SSE, reduced	F^*	P ($F > F^*$)
Full	0.9111×10^9	143			
Reduced	1.05×10^9	145	68849507	10.81	0.00043

Full model: Quadratic model (2); reduced model: change date is 01 April 2020

Model	SSE	DF	SSE, reduced	F^*	P ($F > F^*$)
Full	0.9111×10^9	143			
Reduced	0.9893×10^9	144	78144937	12.26	0.00061

Full model: Quadratic model (3); reduced model: Quadratic model (4)

Model	SSE	DF	SSE, reduced	F^*	P ($F > F^*$)
Full	0.3793×10^9	136			
Reduced	0.5202×10^9	141	140938174	10.10	0.0000

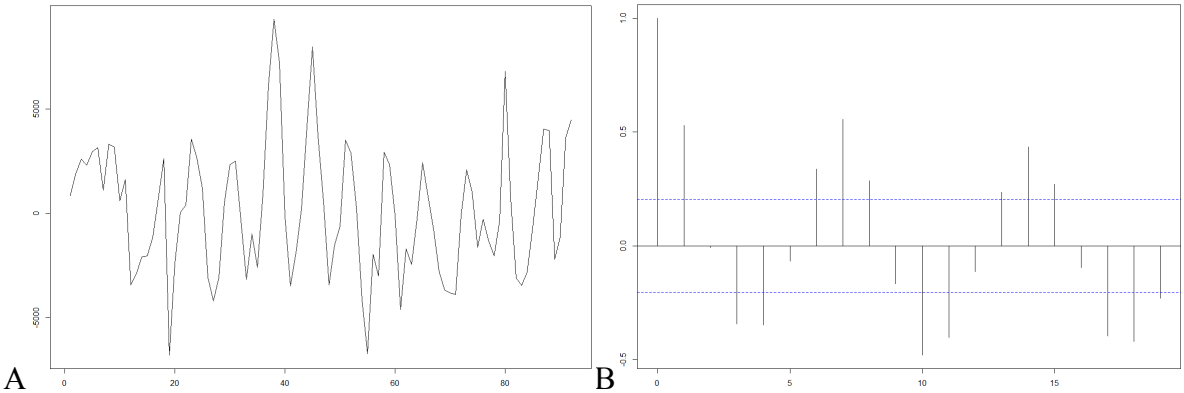


Figure 2: The residual plot (A) and ACF plot (B).

Table 5: The possible change points and corresponding SSE from model (3)

Date	Value of δ	SSE$\times 10^{-8}$
2020-04-01	188425	4.348
2020-04-02	215176	4.102
2020-04-03	244636	3.899
2020-04-04	**277279**	**3.793**
2020-04-05	312519	3.801
2020-04-06	337984	3.885
2020-04-07	367599	4.023
2020-04-08	399388	4.187

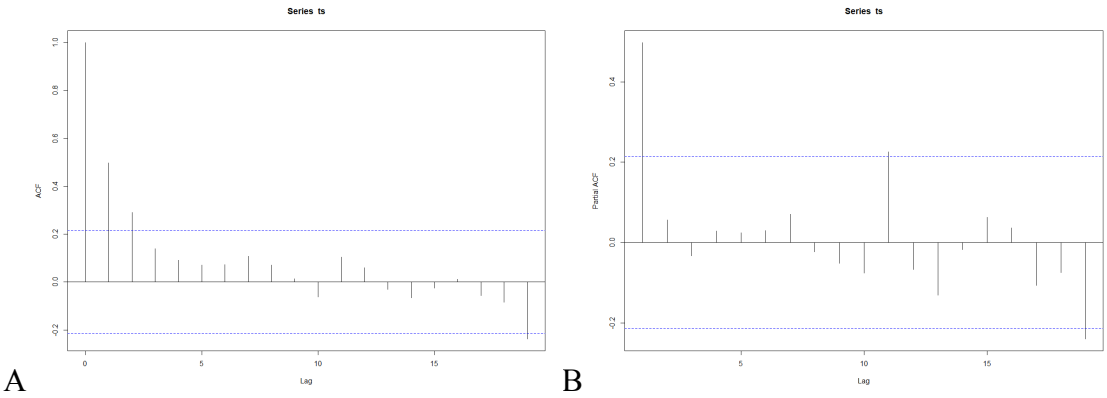
To address the weekly effect, we include the weekday-indicator in the model for the data collected after 27 March 2020:

$$E(y_i) = \beta_0 + \beta_1 x_i + \beta_{12} x_i^2 + \beta_2 (x_i - \delta)_+ + \beta_{22} (x_i - \delta)_+^2 + \left(\sum_{j=1}^{7} \beta_{3j} \cdot \mathbf{1}\{\text{Weekday}_i = j\}\right) \cdot \mathbf{1}\{Date_i \geq 27\ March\ 2020\} \quad (3)$$

Another model for the weekly effect to use the periodical sine and cosine functions:

$$E(y_i) = \beta_0 + \beta_1 x_i + \beta_{12} x_i^2 + \beta_2 (x_i - \delta)_+ + \beta_{22} (x_i - \delta)_+^2 + \left(\beta_s \sin\left(2\pi \cdot \frac{\text{Weekday}_i}{7}\right) + \beta_c \cos\left(2\pi \cdot \frac{\text{Weekday}_i}{7}\right)\right) \cdot \mathbf{1}\{Date_i \geq 27\ March\ 2020\} \quad (4)$$

It can be seen that that model (4) is a reduced model of model (3): $\beta_{3j} = \beta_s \sin\left(\frac{2j\pi}{7}\right) + \beta_c \cos\left(\frac{2j\pi}{7}\right), j = 1, \ldots, 7$. Thus, we can use the ANOVA to test if model (4) is sufficient. The ANOVA test result is listed in Table 4, and the conclusion is that the full model (3) should be used. For the model (3), the possible change point and the corresponding SSE is listed in Table 5, and the result still showed that 04 April 2020 was the change point.



A B

Figure 3: The ACF plot (A) and PACF plot (B) for residue from model (3).

The data was fitted to the weekly effect model (3), and the ACF and PACF plot of resulted residuals were shown in Fig. 3. The PACF plot indicated that the residuals had auto-regression pattern {AR (1)}. The data was then fitted with the weekly-effect model with AR (1). The results are listed in Table 2 and shall be discussed in next section.

The data from individual state was fitted using the following model:

$$E(y_i) = \beta_0 + \beta_1 x_i + \beta_2 (x_i - \delta)_+ + \beta_{22} (x_i - \delta)_+^2 \tag{5}$$

In this model, we removed the second order term before the change point and our motivation was that this model is more sensitive to the change point based on our theoretical study. The results are listed in Table 6 and shall be discussed in next section.

Table 6: The results of individual state data

Model (5): $E(y_i) = \beta_0 + \beta_1 x_i + \beta_2 (x_i - \delta)_+ + \beta_{22} (x_i - \delta)_+^2$

The average delay-time between the date to declare state emergency and change point was 21.8 days with the standard deviation of 5.2 days.

State Name	Change point	$\hat{\beta}_1$	Date to declare state emergence[1]
Alabama	4/2/2020	0.133	3/13/2020
Arizona	3/28/2020	0.229	3/11/2020
California	3/30/2020	0.161	3/4/2020
Colorado	3/25/2020	0.256	3/10/2020
Connecticut	4/5/2020	0.151	3/10/2020
Delaware	4/5/2020	0.166	3/12/2020
D. C.	3/31/2020	0.175	3/11/2020
Florida	4/2/2020	0.156	3/1/2020
Illinois	3/25/2020	0.296	3/9/2020
Indiana	3/30/2020	0.220	3/6/2020
Iowa	4/8/2020	0.168	3/9/2020
Kansas	3/26/2020	0.289	3/9/2020
Kentucky	4/6/2020	0.119	3/9/2020
Louisiana	4/1/2020	0.276	3/11/2020
Maine	3/27/2020	0.106	3/15/2020

Maryland	4/2/2020	0.186	3/5/2020
Massachusetts	3/27/2020	0.328	3/10/2020
Michigan	3/31/2020	0.205	3/11/2020
Mississippi	4/2/2020	0.115	3/4/2020
Missouri	3/31/2020	0.185	3/13/2020
Nebraska	4/8/2020	0.156	3/13/2020
Nevada	3/28/2020	0.221	3/12/2020
New Hampshire	3/28/2020	0.173	3/13/2020
New Jersey	3/29/2020	0.237	3/9/2020
New Mexico	4/6/2020	0.120	3/11/2020
New York	3/22/2020	0.436	3/7/2020
North Carolina	3/26/2020	0.243	3/10/2020
Ohio	4/2/2020	0.144	3/9/2020
Pennsylvania	4/3/2020	0.187	3/6/2020
Rhode Island	4/8/2020	0.157	3/9/2020
South Carolina	3/31/2020	0.168	3/13/2020
South Dakota	4/8/2020	0.254	3/13/2020
Tennessee	3/30/2020	0.143	3/12/2020
Texas	4/5/2020	0.141	3/13/2020
Utah	3/27/2020	0.229	3/6/2020
Virginia	3/31/2020	0.189	3/12/2020
Washington	3/26/2020	0.159	2/29/2020

[1]*: the date of the declaration of state emergency is from wikipedia.org (https://en.wikipedia.org/wiki/U.S._state_and_local_government_response_to_the_COVID-19_pandemic)*

3. Results and Discussions

For the US data, from Table 2 the fitted equation is given by

$$E(y) = 132 + 0.287x - 5.143 \times 10^{-7}x^2 - 0.01278(x - 277279)_+ + 5.170 \times 10^{-7}(x - 277279)_+^2 + \text{weekly effect}$$

$$= \begin{cases} 132 + 0.287x - 5.143 \times 10^{-7}x^2 + \text{weekly effect}, & \text{before 04 April 2020} \\ 43424 - 0.01248x + 2.7 \times 10^{-9}x^2 + \text{weekly effect}, & \text{after 04 April 2020} \end{cases}$$

In this equation, $\hat{\beta}_1$=0.287 suggested that at the early stage when the total number of confirmed cases was small, each patient had 28.7% chance to infect another healthy person each day. Since a Covid-19 patient usually recovered within 2 weeks, R_0 value can be estimated by $0.287 \times 14 = 4.01$, which was consistent with the published results (median value 5.7 with 95% confidence interval: 3.8 - 8.9, Steven Sanche, *et. al.*, 2020). $\hat{\beta}_{12}$<0 indicated that even before the change point, the rate was decreasing from 0.287. In fact, the rate at 04 April 2020 can be calculated as $0.287 - 5.143 \times 10^{-7} \times 277279 = 0.144$, which was only half of the original rate. In our study of the State data, we found that several states had their change points in late March. This could be the due to the reason that the rate was decreased to 0.144, as several states had already slowed down.

The change point was 04 April 2020 for the whole US data. Because the median incubation time of Covid-19 was 4-5 days, implying that what resulted in the change point should have played the role at least one week before 04 April 2020. This seems to indicate that

the change point could be resulted from the issuance of National Emergency on 13 March 2020. If it was true, it suggested that the effect of people's behavior would be reflected by the change point about 21 days later. The same lag effect was also observed at the state level.

Listed in Table 5 are the fitting results for the data from individual states. Washington State was the first with the outbreak of Covid-19. The scatter plot of the data superimposed with the fitted curve is shown in Figure 4A. Before the change point, the daily new cases were increasing. After 26 March 2020, however, the number started to decrease. But the number seemed to comeback recently. The estimate is $\hat{\beta}_1$=0.154, which indicated that the initial rate in Washington States was less than the average rate (0.287) of the US. The state emergency was declared on 29 February 2020, and the change point was on 26 March 2020, thus it showed about 25-day delay.

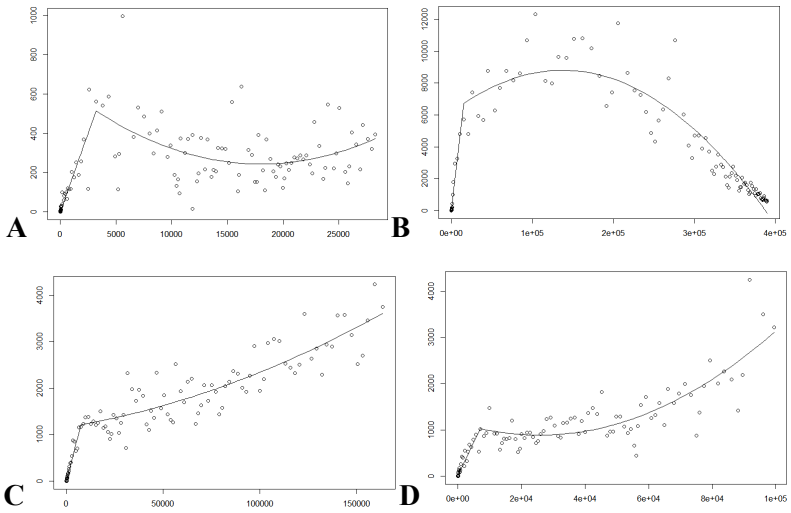


Figure 4: The scatter plot of state data. The *y*-axis is the number of daily new cases, and the *x*-axis is the total number of cases up to the previous day. **A**: Washington State. **B**: New York State. **C**: California State. **D**: Texas State.

New York State was a hot spot in March. The plot is shown in Figure 4B. The estimate $\hat{\beta}_1$ is equal to 0.436, which is the highest among all states. The high rate could be due to its high population density and heavy public transportation. The state emergency was declared on 07 March 2020 and the change point was 22 March 2020, which lagged behind 15 days. After 22 March 2020, the daily new cases stayed with high value and then dropped down. This seemed to indicate that the Covid-19 appeared to be controlled.

The plot of data from California State is shown in Figure 4C. The state emergency was declared on 04 March 2020 and the change point was 30 March 2020, which lagged behind about 26 days. However, after the change point, the daily new cases were only slowing down and still kept increasing. To further control Covid-19, more efforts will be needed. The plot of data from Texas is shown in Figure 4D. The state emergency was declared on 13 March 2020 and the change point was 05 April 2020. For Texas, the lag time was 22 days.

As we discussed before, the estimate $\hat{\beta}_1$ for each state was proportional to R_0 for that state before any prevention measures were used. Some states, similar to New York State, like Massachusetts and Illinois, have big metropolitan areas (Boston in MA, and Chicago in IL) with high population density and heavily public transportation. Thus, the estimate $\hat{\beta}_1$ of these

states were relatively higher than the rest. Other states, like Mississippi and New Mexico, have no such big cities, and usually had lower estimate $\hat{\beta}_1$.

Overall, the data from most states showed a change point pattern. Before the point, the daily new cases were proportional to the total cases, similar to the whole US data. By comparing the change point and the date when the state emergency was declared in Table 6, we found that the average delay-period is 21.8 days. This suggested that if there is another change point, what happen 3 weeks before would likely be the causes of the change.

4. Consistency and Asymptotic Normality

Here, we present consistency and asymptotic normality results and omit the proofs. What is novel here is that we model the change point and the possible non-linear relationship simultaneously, whereas a typical change point model involves in only $(x-\delta)_+$. This is *a continuous second-order free spline model with one knot*.

To prove asymptotic normality, we have to deal with the irregular criterion function $\left(y_i - \boldsymbol{\beta}^T \boldsymbol{z_i}(\delta)\right)^2$, in which the truncated power function x_+ is not differentiable. Thanks to Theorem 5.23 of van der Vaart (1998), we have obtained a quick result at the price of boundedness Assumption 4. In other words, with careful elaboration, we believe this assumption (and others as well) can be relaxed to the boundedness assumption of the knot parameter δ as in the case of consistency, see Wu, *et. al.* (2019).

Consider that $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ satisfy the second-order free spline model,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 (x-\delta)_+ + \beta_4 (x-\delta)_+^2 + \epsilon_i,\ i = 1, \dots, n,$$

where $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ are i.i.d. random errors with $E(\epsilon_i) = 0$ and $V(\epsilon_i) = \sigma^2 < \infty$. Here $x_1, x_2, \dots, x_n$ are assumed to be non-random, the coefficients β and the knote δ are unknown parameters to be estimated.

Denote $\boldsymbol{\beta} = (\beta_0,\ \beta_1,\ \beta_2,\ \beta_3,\ \beta_4)^T$, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \delta)^T$, and $\boldsymbol{z_i}(\delta) = (1,\ x_i,\ x_2^2,\ (x-\delta)_+,\ (x-\delta)_+^2)^T$**.** Using these symbols, we can write

$$y_i = \boldsymbol{\beta}^T z_i(\delta) + \epsilon_i, \qquad i = 1, \dots, n.$$

We estimate $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \delta)^T$ by the least squares estimate (LSE) $\widehat{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{\beta}}^T, \hat{\delta}\right)^T$, that is,

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} S_n(\boldsymbol{\theta}), where\ S_n(\boldsymbol{\theta}) = S_n(\boldsymbol{\beta}^T, \delta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{\beta}^T \boldsymbol{z_i}(\delta)\right)^2. \tag{6}$$

For $\delta \in \Delta \subset R$ fixed, the minimization (6) simplifies to the usual LSE problem. Let $\boldsymbol{Z}(\delta)$ be the $n \times 5$ matrix consisting of $\boldsymbol{z_1}(\delta), \boldsymbol{z_2}(\delta), \dots, \boldsymbol{z_n}(\delta)$ as its rows, and let $\boldsymbol{y} = (y_1, y_2, \dots, y_n)^T$. If $\boldsymbol{Z}(\delta)$ has full rank 5, then the LES $\hat{\beta}(\delta)$ is given by

$$\hat{\beta}(\delta) = [\boldsymbol{Z}^T(\delta)\boldsymbol{Z}(\delta)]^{-1}\boldsymbol{Z}^T(\delta)\boldsymbol{y}. \tag{7}$$

As a result, the minimization (6) becomes minimizing the new objective over $\delta \in \Delta$:

$$\hat{\delta} = \arg\min_{\delta \in \Delta} \tilde{S}_n(\delta),\ \tilde{S}_n(\delta) = S_n(\boldsymbol{\beta}^T(\delta), \delta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{\beta}^T(\delta)\boldsymbol{z_i}(\delta)\right)^2.$$

Assumption 1. There exists a compact subset Δ of $\mathbb{R}$ and a matrix function $\boldsymbol{M}(\delta_1, \delta_2)$, $\delta_1, \delta_2 \in \Delta$, such that

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z_i}(\delta_1)\boldsymbol{z}_i(\delta_2)^T \to \boldsymbol{M}(\delta_1, \delta_2),$$

uniformly in $\delta_1, \delta_2 \in \Delta$, that $\boldsymbol{M}(\delta, \delta)$ is positive definite on Δ, and that $\boldsymbol{T}(\delta) = \boldsymbol{M}(\delta_0, \delta_0) - \boldsymbol{M}(\delta_0, \delta)\boldsymbol{M^{-1}}(\delta, \delta)\boldsymbol{M}(\delta_0, \delta)$ has a unique zero solution at $\delta = \delta_0$.

Assumption 2. For all $n \geq 1$, $\sup_{1\leq i\leq n}\{|x_i|\} \leq M_x < \infty$ for some constant M_x.

Note **Assumption 1** ensures that the maximizer is well-separated and unique. It is a typical assumption for establishing consistency of M-estimators, see Chapter 5 of Van der Vaart (1998), Yu and Ruppert (2002) and Wu, *et al.* (2019).

Theorem 1. Assume Assumptions 1 and 2. Then the LSE $\widehat{\boldsymbol{\theta}}$ converges in probability the true value $\boldsymbol{\theta_0} = (\boldsymbol{\beta_0^T}, \delta_0)$ of parameter, *i.e.*,$\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta_0}$, in probability.

Remark. If X_i are random, consistency still holds provided that ϵ_i and X_i are independent with $EX^2 < \infty$, and the convergence in **Assumption 1** is modified to convergence in probability.

We need the following assumptions to assure asymptotic normality.

Assumption 3. $X_1, X_2, \ldots, X_n$ are i.i.d. with a common continuous density function f, X_i and ϵ_i are independent for all i, and $E(X^8) < \infty$.

Because x_+ is not differentiable, asymptotic normality was proved using the empirical process theory. This requires the square-integrability of the envelope function, which is a polynomial of x of fourth degree, leading to finite 8th moment assumption.

Assumption 4. There exists a neighborhood of $\boldsymbol{\theta_0}$, such that $\forall \boldsymbol{\theta} \in N(\boldsymbol{\theta_0})$, $\|\boldsymbol{\theta}\| \leq B_0 < \infty$ for some $B_0 > 0$**.**

Let $\mu(\boldsymbol{\theta}) = E\big(\boldsymbol{S}_n(\boldsymbol{\beta}^T, \delta)\big)$, $\dot{\mu}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}}\mu(\boldsymbol{\theta})$ be the 6-dimensioinal derivative vector and $\boldsymbol{V}(\boldsymbol{\theta_0}) = \ddot{\mu}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\mu(\boldsymbol{\theta})$ be the 6-by-6 matrix of second partial derivatives.

Assumption 5. $\dot{\mu}(\boldsymbol{\theta_0}) = \boldsymbol{0}$ and the matrix $V(\boldsymbol{\theta_0})$ is nonsingular.

Theorem 2. Assume Assumptions 3-5. If the LSE is consistent, *i.e.*, $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} \to \boldsymbol{\theta_0}$ in probability, then $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ is asymptotically linear,

$$\sqrt{n}\big(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta_0}\big) = -\boldsymbol{V^{-1}}(\boldsymbol{\theta_0})\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \dot{m}_{\boldsymbol{\theta_0}}(X_i, Y_i) + o_p(1)$$

where $\dot{\boldsymbol{m}}_{\theta}(x, y) = \frac{\partial}{\partial(\boldsymbol{\beta}^T, \delta)^T}\big(y - \boldsymbol{\beta}^T\boldsymbol{z}(\delta)\big)^2$. Hence,

$$\sqrt{n}\big(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta_0}\big) \sim N\Big(0, \boldsymbol{V^{-1}}(\boldsymbol{\theta_0})E\big[\dot{\boldsymbol{m}}_{\boldsymbol{\theta_0}}(X_1, Y_1)\dot{\boldsymbol{m}}_{\boldsymbol{\theta_0}}(X_1, Y_1)^T\big]\boldsymbol{V}^{-1}(\boldsymbol{\theta_0})\Big).$$

5. Future study

During the preparation of this paper, we have noticed that there was a second outbreak in the US at the end of June. Our approach can be easily generalized to multiple change points. Currently, we work on the theoretical development in the framework of time series model with multiple change points.

Acknowledgements

The authors are indeed thankful to Prof. Jyoti Sarkar and Prof. Vinod Kumar Gupta, Chair Editor, for suggestions and encouragement.

References

Bai, J. (1999). Estimation of a change point in multiple regression models. *The Review of Economics and Statistics,* **79(4)**, 551-563.

Julious, S. (2001). Inference and estimation in a changepoint regression problem. *The Statistician*, **50**, 51-61.

Sanche, S., Lin, Y., Xu, C., Romero-Severson, E., Hengartner, N. and Ke, R. (2020). High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerging Infectious Diseases*, **26(7)**, 1470-1477.

Van der Vaart, A. W. (1998). *M- and Z- Estimators*. *Asymptotic Statistics*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511802256.006.

Wu, J., Peng, H. and Tu, W. (2019). Large-sample estimation and inference in multivariate single-index models. *Journal of Multivariate Analysis*, **171**, 382-396.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models, *Journal of American Statistical Association,* **97,** 1042–1054.